地域情報抽出技術　Twitter　自然言語処理

Technology Reports

Twitterを用いた地域イベント発見技術

Twitterはリアルタイムに情報が共有されるソーシャルネットワークサービスであり，大量のさまざまな話題の情報がリアルタイムで共有されている．本稿では自然言語処理技術を活用して，日本全国の各地域で開催されるイベント情報をTwitterから自動で発見する技術について解説した後，本技術を活用したサービス「街のイベント情報」を紹介する．

先進技術研究所　山田（やまだ）　渉（わたる）
サービスイノベーション部　落合（おちあい）　桂一（けいいち）　菊地（きくち）　悠（はるか）

1. まえがき

スマートフォンの普及に伴って，さまざまな種類の位置情報サービスが多数提供されている．例えばTrip Advisor®*1は世界中のホテル・レストラン情報やその評判情報を，ぐるたび®*2は日本各地の名物グルメ情報を提供している．このような位置情報サービスをさらに魅力あるものにするためには，観光スポットやイベント，名産品などの地域情報について，最新の情報をできる限り多く提供することが重要である．

しかし，最新の地域情報の提供を維持するためには多大な労力がかかる．さらに地域情報の中でもイベント情報は，日々新たなものが開催されるため，情報の鮮度を確保するには特に頻繁な更新が必要となるが，手動での対応には限界がある．

これを解決するためTwitter*3を活用して日本全国各地で開催されるイベント情報を自動で発見する技術を開発した．Twitterとはソーシャルネットワークサービスの1種であり，140文字のテキスト（ツイート*4）を投稿・共有できるサービスである．Twitterでは，ユーザの日常生活で起きたことや新製品，ニュース，そしてイベント告知などの大量のさまざまな話題の情報がリアルタイムで共有されている．特にTwitterを用いたイベントの告知は誰でも手軽に行うことができるため，花火大会や地域のお祭りといった公共性の高いイベントに限らず，店舗のフェアやインディーズバンドのライブといったさまざまな種類のイベント情報が大量に告知されている．

本技術は，自然言語処理技術*5を活用してツイートからイベント情報を自動で発見するもので，大量のツイートからイベント情報の有無だけでなく，そのイベントの名称，開催場所，開催期間の3つ組を8〜9割程度の精度で抽出できる．

本稿では，この地域イベント情報自動発見技術とともに，本技術を適用した「街のイベント情報」サービスについて解説する．

2. 地域イベント発見技術概要

地域イベント発見技術の概要について，**図1**に示すデモンストレーションアプリの動作画面を用いて解説する．これは本技術を可視化し検証するために用意したものであり，自動抽出されたイベント情報を，開催期間，開催場所に応じて表示しており，ユーザは各地域のイベント情報を簡単に発見することができる．

*1 **Trip Advisor®**：各国のレストランやホテル情報，口コミ情報を提供するサービス．TripAdvisor, LLCの商標または登録商標．
*2 **ぐるたび®**：日本各地の名産品を紹介するサービス．株式会社ぐるなびの登録商標．
*3 **Twitter**：アメリカ合衆国また他の国々におけるTwitter, Inc.の登録商標．
*4 **ツイート**：Twitter社の提供するマイクロブログサービスにおける記事のこと．
*5 **自然言語処理技術**：人間が日常的に使っている言語（自然言語）をコンピュータに処理させる技術．

※ツイート文は架空の例示であり実在のものではありません．

図1　イベント発見技術のプロトタイプイメージ

また図1には，自動抽出された各地域のイベント情報のうち，2015年9月4日（金）に開催されると推定されたものを表示している．各イベント情報はイベントの名称，開催場所，開催期間の3つ組とそれらの抽出元のツイート情報が格納されている．例えば，図1中の吹き出しで示している“ナイトアクアリウム”の例では，“ナイトアクアリウム”というイベントの名称，“新江ノ島水族館”というイベントの開催場所，“2015年7月20日～2015年11月30日”の開催期間の3つ組が含まれている．またイベント名称を含むツイートを再検索することで，他のユーザのイベントに関するツイートも見ることができる．図1の地図上の数字は，“ナイトアクアリウム”のように自動抽出された各地域のイベント情報の合計件数である．例えば，9月14日に東京近郊で536件のイベントが開催されることを示している．

イベントの発見件数は，時期や休日か平日かで変動するが，新規イベントを全国で毎日150件程度，ひと月あたり4,000～5,000件程度抽出することができる．この件数はイベント情報のデータベースとしては国内最大規模である．

従来は，位置情報付きツイートの投稿数が急上昇している場所を特定し，イベントを抽出する方法が主だった[1] [2]．これらの手法では，イベントの発生によって周辺地域の投稿数が変化することを前提にしているため，地震の発生や有名アーティストのライブなど投稿数が多い大規模なものが対象となり，投稿数が比較的少ない小規模なイベントは抽出することができない．またイベントが発生する前にイベント情報を抽出することはできない．

本技術では，位置情報付きツイートの投稿件数ではなく，イベント情報を含むツイートが持つ自然言語的な特徴に着目し，機械学習*6技術を用いてイベント情報を抽出する．そのためイベントの告知ツイートが1件だけであったとしても抽出対象となるため，大規模なイベントだけでなく，地域で開催されるような小規模イベント情報も発見することが可能である．

3. システム構成および処理の流れ

地域イベント発見技術は，**図2**に示すように地名抽出部とイベント情報抽出部から構成されている．それ

＊6　**機械学習**：事例をもとにした統計処理により，計算機に入力と出力の関係を学習させる枠組み．

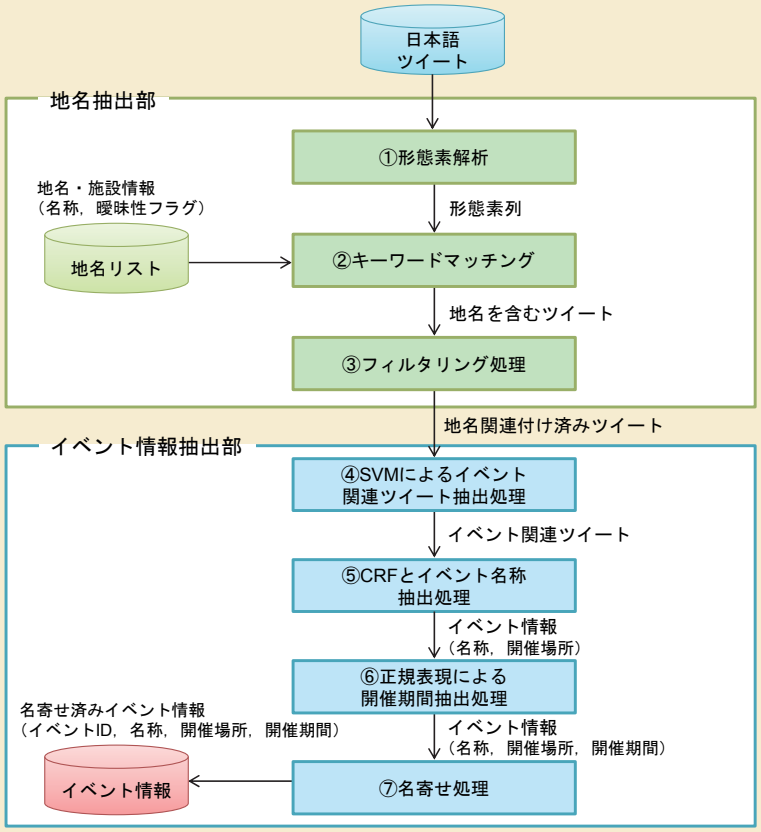

図2　地域イベント発見技術の全体処理概要

ぞれの処理内容を解説する．

3.1　地名抽出部

地名抽出部は，日本語のツイートを対象として，ツイートがどの場所について言及しているのかを解析して，地名とツイートの関連付けを行う．この関連付けは図2①～③までの3ステップで構成されている．まず日本語のツイートを対象として，形態素解析*7を実行する（図2①）．次にあらかじめ用意した，地域や施設の名称と後述する曖昧性を示すフラグなどを含んだ地名リストを参照し，地名と一致する名詞が本文中に含まれるツイートを抽出する（図2②）．最後に曖昧性がある地名を含むツイートに対してフィルタリングを行う（図2③）．曖昧性のある地名とは，人名としても用いられるものや，同名で異なる場所が複数存在するため特定の場所を一意に示すとは限らない地名のことである．その例としては，苗字の“松島”と宮城県の観光名所の“松島”がある．同名の地名が存在する例としては，京都府の“円山公園”と北海道の“円山公園”などが挙げられる．

この曖昧性を除去し，正しくフィルタリングを行うために共起語と機械学習を用いた処理を行う．このフィルタリングの詳細については文献[3]を参照されたい．

以上の処理によって，地名抽出部は日本語ツイートを地名と関連付け，曖昧性の除去を行い，イベント情報抽出部へと出力する．

3.2　イベント情報抽出部

イベント情報抽出部では，地名関連付け済みツイートから，イベント名称と開催期間の抽出を行う．本処理は大きく4つのステップで構成される．第一に，地名と関連付け済みのツイートからイベント名称や開催

*7　**形態素解析**：文章を形態素と呼ばれる意味のある単語の最小単位に区切る技術のこと．

期間といった情報を含むツイートを抽出するイベント関連ツイート抽出処理を行う（図2④）．第二に，イベント情報を含むツイートからイベント名称の抽出を行う．また開催場所の情報には地名抽出部を用いてツイートと関連付けられた地名を割り当てる（図2⑤）．第三に，イベント名称が抽出されたツイートから開催期間を抽出する（図2⑥）．第四に抽出されたイベント情報に対して，開催場所とイベント名称の類似度を用いて同一のものかを判定する名寄せ処理を行い，イベント情報ごとにIDを付与する（図2⑦）．それぞれの処理の詳細を以下に解説する．

(1)イベント関連ツイート抽出処理

イベント関連ツイート抽出処理では，地名関連付け済みツイートから機械学習を用いてイベントに関係するツイートのみを抽出する．本処理では，イベントに関するツイートは“開催”や“祭り”といった語が文中に出現しやすいといった特徴を分類器*8に学習させて，イベントに関連したツイートを地名関連付け済みツイートから抽出する．大量のツイートを高速に処理するため，線形SVM（Support Vector Machine）[4]と呼ばれるアルゴリズムを採用している．このアルゴリズムは**図3**のように，2つのフェーズに分かれている．

(a)学習フェーズではまずツイートを収集して，各ツイートを目視で確認して，イベントに関係したツイートとイベントに関係のないツイートのどちらかのラベルを付与する．次に，ラベルを付与したツイートから素性（そせい）*9を抽出する．素性には各単語がそれぞれ何回出現したかという単語の出現回数を用いている．そのため“開催”や“展示”など，イベントに関連するツイートに含まれやすく，かつイベントに関係しないツイートには含まれにくい単語が多く出現するツイートほど，イベントに関連していると判定される．

(b)推定フェーズでは，学習フェーズで構築した分類器を用いて，地名関連付け済みの各ツイートがイベント情報を含むかどうかを判別する．そしてイベント情報を含むと判別されたツイートはイベント名称抽出処理と出力される．

(2)イベント名称抽出処理

イベント名称抽出処理は，SVMによってイベントと関連すると判別されたツイートの本文に含まれるイベント名称を抽出する．本処理では

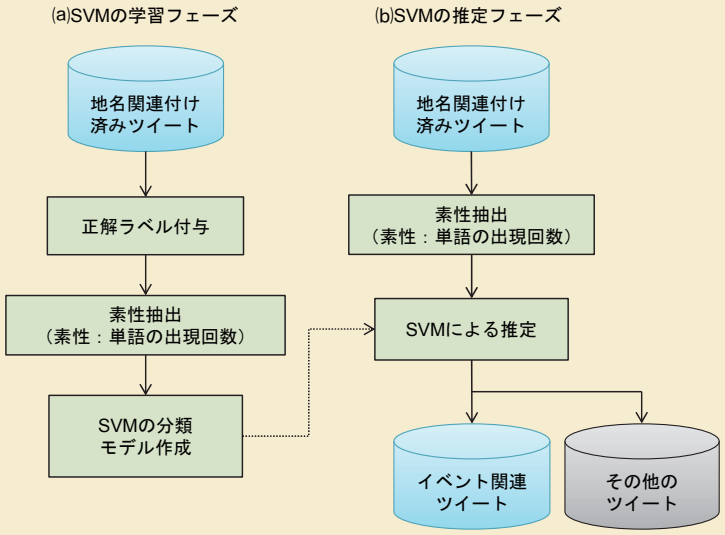

図3　SVMによるイベント関連ツイートの抽出処理

*8 **分類器**：入力をその特徴量に基づいてあらかじめ定められた分類先のいずれかに分類する装置．
*9 **素性**：自然言語処理における特徴量のこと．

CRF（Conditional Random Fields）*10[4]という機械学習のアルゴリズムを用いて，イベント名称を抽出する．

CRFによるイベント名称の抽出も**図4**のように2つのフェーズに分かれている．

(a)学習フェーズでは，イベントに関連するツイートに対して，どの部分がイベント名称に相当するか，またどの部分がイベント名称に関係ないかといったラベルを付与する．そしてCRFは付与されたラベルから，イベント名称には“祭り”という表現が含まれやすい，“October fest”のようにアルファベットが続きやすいといったさまざまな特徴を学習し，分類モデルを作成する．また素性には各単語の読みや表記，品詞，文字数などを使用している．

(b)推定フェーズでは，学習フェーズで作成した分類モデルを用いて，イベントに関連していると判定されたツイートのうち，どの部分がイベントの名称に関連しているかを判別する．そしてイベント名称が抽出されたツイートは次の開催期間抽出に出力される．

(3)開催期間抽出処理

①正規表現を用いた抽出

開催期間抽出処理では，正規表現を用いて，ツイートの本文中に含まれるイベントの開催期間を抽出する．正規表現とは自然言語処理の手法の1つで，あらかじめ定義されたパターンに文字列が該当するかを判定したり，そのパターンに該当する文字列を文中から抽出したりすることができる．例えば「¥d{2,4}年¥d{1,2}月¥d{1,2}日」というパターンを定義することで「2016年1月1日」や「16年12月31日」などの2桁から4桁の年と1桁から2桁の月と日から構成される文字列を抽出することができる．開催期間抽出では，日付に関連するパターンをあらかじめ大量に登録しておき，それらをパターンマッチングに用いて文中に含まれる日付を抽出する．

②日付の補完

抽出した日付は「1月1日」や「本日」などのように必ずしも年月日が全て含まれているとは限らない．これらについてはツイートの投稿日付を参照して，適切な年月日を補完する．また「2016年1月1日から3日まで開催」といった，開催日が単一の日付ではなく，期間となっている場合に対応するため，抽出した日付間に「より」や「から」，「～」などの期間であることを

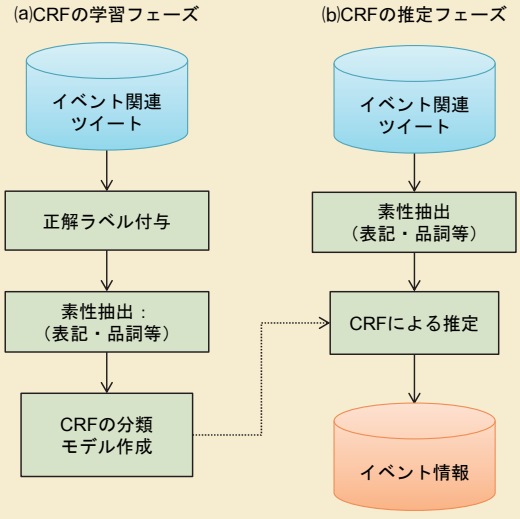

図4　CRFによるイベント名称抽出処理

*10 **CRF**：条件付確率場．入力された要素が連なったもの（系列）に対して，その特徴量に基づいてあらかじめ定められたラベルを付与する手法の一種．

示す語が，抽出した日付の間にあるかを確認する．そして，このような期間である語が入っている場合には，さらに日付の前後関係などを確認し，開催期間として抽出する．

以上の処理によってイベントの名称と開催期間，また地名抽出部によって割り当てられた開催場所の3つ組を抽出する．

⑷名寄せ処理

開催期間抽出までの一連の処理によって，イベント情報は抽出できるものの，1つのイベントに対して複数のツイートが告知を行っている場合があるため，抽出されたイベント情報には重複するものが存在する．

しかもユーザによって，「21世紀の未来展」と「21世紀のみらい展」のように，イベント名称の表記が異なる場合があるため，異名で重複したイベント情報が含まれている．

そこで本処理ではイベントの名称と開催場所の2つの情報を用いてイベント情報の同一性の判定を行ったうえで，イベントIDを割り当てる．これはイベント情報を管理するための番号であり，重複するイベント情報には同一のイベントIDが割り当てられる．

本処理では**図5**のように，抽出したイベント情報を開催場所ごとにグルーピングをする．次に，開催場所が同じイベント情報どうしのペアを全通り作成する．そして作成したペアのイベント名称の類似度を計算し，しきい値以上であれば，同一のイベント情報と判定し，それぞれに同一のイベントIDを割り当てる．イベント名称の類似度には最長共通部分列比という類似度を用いている．本処理により，図5のように同じ場所で開催される類似した名前のイベント情報をまとめることができる．

4. 「街のイベント情報」サービス

本技術を活用した「街のイベント情報」サービスが2015年5月12日より，dメニューのリアルタイム検索コーナー内で提供されている．本サービスは，**図6**に示す通り，本技術で収集したイベント情報のうち，

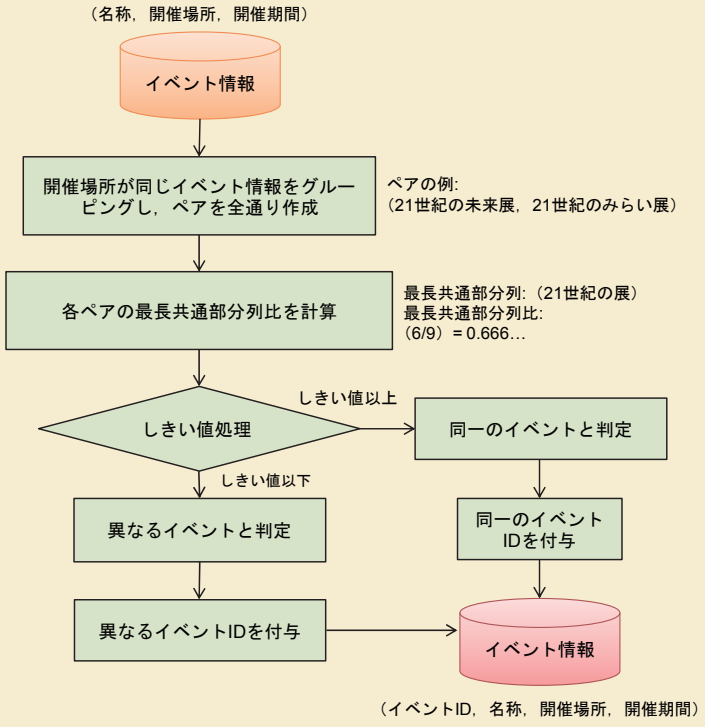

図5　イベント情報の名寄せ処理

図6 「街のイベント情報」サービス画面

ユーザの現在地付近のものを表示することで，近くの話題のイベントを見つけることを可能とするものである．現在地のイベントだけでなく地図やカレンダーで場所と日付を指定してイベント情報を調べることができるため，旅行やおでかけの計画に利用することもできる．

5. あとがき

本稿では，ツイートを情報源とした地域イベント情報の自動発見技術の概要を解説し，本技術を活用した「街のイベント情報」サービスを紹介した．今後は本技術を活用して全国各地のイベント情報をコンテンツ化しそれによって地方創生につながるサービスを創造していく．また各地の名産品やその評判といったイベント情報以外の地域情報抽出技術の研究開発を行う．

文　献

[1] L. Chen and A. Roy: “Event Detection from Flickr Data through Wavelet-based Spatial Analysis,” Proc. of the 18th ACM Conference on Information and Knowledge Management, 2009.

[2] R. Lee and K. Sumiya: “Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection,” Proc. of 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, pp.1-10, 2010.

[3] 落合，ほか：“位置に関連するツイート解析技術とその応用，”本誌，Vol.22, No.2，pp.30-35，Jul. 2014.

[4] Corinna. C and Vladimir. V: “Support-Vector Networks, Machine Learning,” Vol.20, pp.273-297, 1995.

[5] L. John, M. Andrew and C. Feramdo: “Conditional randomfields: Probablistic models for segmenting and labeling sequence data,” Proc. of the Eighteen International Conference on Machine Learning, pp.282-289, 2001.